Regression in the Presence of Strongly Correlated Data

Christopher S. Jones and John M. Finn, T-15; and Nicolas Hengartner, D-1

In plasma equilibrium reconstruction in plasmas one attempts to find equilibrium parameters that give the best fit to a set of experimental measurements, such as the magnetic field at the boundary or internal pressure measurements. This reconstruction problem is patently nonlinear in its parameters because of the nonlinearity in the magnetohydrodynamic (MHD) equilibrium equation, the Grad-Shafranov equation. We have studied the effect of correlations on the equilibrium reconstruction problem [1]. These correlated errors represent fluctuations in the data due to shorter time scale physical processes in the plasma, and were treated as correlated gaussian noise. We observed that the fit becomes perfect in the limit of large correlations. The quality of the fit was measured by the covariance matrix of the parameters, or in Bayesian language, the *posterior covariance matrix* $\mathbf{C}_p$. The case studied in [1] had noise characterized by a single parameter, the *correlation length* δ. In terms of δ, Cp was, quite surprisingly, found to go to zero in the limit of perfect correlation $\delta \to \infty$. Thus, if the physical processes that are modeled as noise are highly correlated in space, the estimate may be much better than expected.

More recently, we have performed analysis to understand better this result and determine its range of validity. We have found that it occurs as well in linear estimation, i.e., linear regression, and we have studied the linear version because the effect is most transparent there. We have found that if the data can indeed be described in terms of a single parameter δ, the best linear unbiased estimate in the limit $\delta \to \infty$ generically gives $\mathbf{C}_p \to 0$.

The simplest form of this effect occurs when measuring a single variable a with repeated correlated measurements. That is, we have, for $i = 1, \ldots, n$,

$$y_i = a + \eta i;$$

the noise η_i has zero mean and a data covariance matrix $C_{ij} = \sigma_i \sigma_j R_{ij}$, with $R_{ij} = \varepsilon^{|i-j|}$. In terms of the correlation length, the correlation parameter ε equals $\varepsilon^{-\Delta/\delta}$, where Δ is the distance (in space or time) between successive measurements. The parameters σ_i are the standard deviations of the individual errors and R_{ij} is the matrix of correlations.

These data are of the form

$$\mathbf{y} = \mathbf{X}\mathbf{a} + \eta,$$

where **X** is the design matrix and $\mathbf{a} = (a_l, \ldots, a_m$ for $m < n$ is the estimate. For (Eq. 1), i.e., $m = 1$, $n = 2$, and $\mathbf{X} = (1, 1)^T$, we recover the familiar weighted least squares result for $\varepsilon = 0$

$$a_e = \frac{y_1/\sigma_1^2 + y_2/\sigma_2^2}{1/\sigma_1^2 + 1/\sigma_2^2}; \quad V_p = \frac{1}{1/\sigma_1^2 + 1/\sigma_2^2},$$

where V_p is the posterior variance. For the generic case $\sigma_1 \neq \sigma_2$ and $\varepsilon \to 1$, we have

$$a_e \to \frac{y_1/\sigma_1 - y_2/\sigma_2}{1/\sigma_1 - 1/\sigma_2}; \quad V_p \to \frac{2\,(1-\epsilon)}{(1/\sigma_1 - 1/\sigma_2)^2}.$$

Note that $V_p \to 0$ as $\varepsilon \to 1$. Also, if we assume $\sigma_2 > \sigma_1$ without loss of generality, then y_2 has *negative weighting.* That is, the estimate a_e is outside the range of the data (y_1, y_2). The exception to this rule is when $\sigma_1 = \sigma_2$. The posterior variance V_p as a function of the correlation ε is shown in Fig. 1; for all cases except $\sigma_2 = \sigma_1$, V_p increases for small ε and decreases to zero as $\varepsilon \to 1$. For this case, negative weighting occurs for ε to the right of the peak of V_p.

Negative weighting has been observed in the nuclear data community, where it is discussed — but without mention of the result $V_p \rightarrow 0$ — in the context of *Peelle's pertinent puzzle* [2, 3].

The generalization of this result to arbitrary *m*, *n* is that the *m x m* posterior covariance matrix $\mathbf{C}_p$ goes to zero except for the special cases when the vector of variations $(\sigma_1, \ldots, \sigma_n)$ is in the range space of the design matrix **X**. Since this possibility is measure zero, the result $\mathbf{C}_p \rightarrow 0$ holds generically.

For more information contact Christopher S. Jones at csjones@lanl.gov.

[1] C.S. Jones and J.M. Finn, *Nucl. Fusion* **46**, 335–349 (2006).
[2] G. D'Agostini, *Nucl. Instr. Meth. Phys. Res.* **346**, 306 (1994).
[3] K.M. Hanson, et al., "Probabilistic Interpretation of Peelle's Pertinent Puzzle and Its Resolution," *Proc. Int. Conf. Nuc. Data Sci. and Tech.*, R.C. Haight, et al., Eds., *AIP Conf. Proc.* **769**, 304–307 (AIP, Melville, 2005).

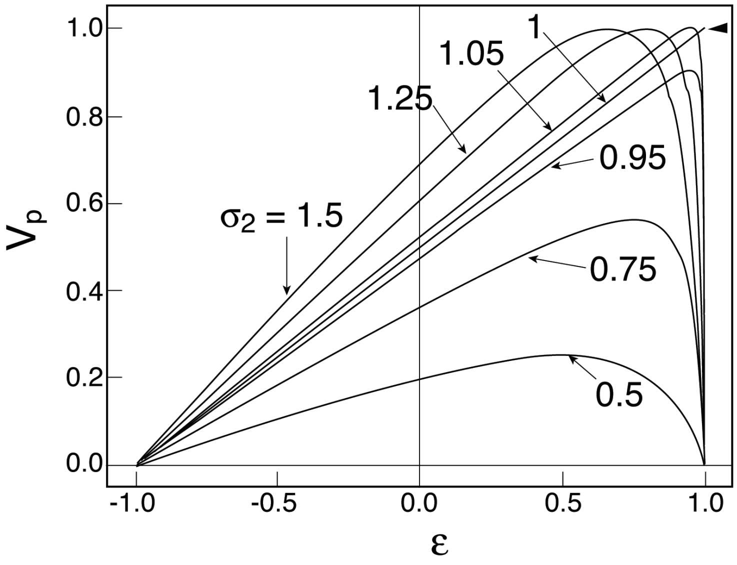

Fig. 1.
Variance V_p as a function of ε for 7 values of σ_2 (0.5, 0.75, 0.95, 1, 1.05, 1.25, 1.5) and $\sigma_1 = 1$. For $\sigma_2 \neq \sigma_1$, $V_p \rightarrow 0$ as $\varepsilon \rightarrow 1$.

Los Alamos
NATIONAL LABORATORY
EST. 1943